Development of Tools for Analyzing and Sharing Proteomics Data

Harald Barsnes

Dissertation for the degree philosophiae doctor (PhD)

UNIVERSITY OF BERGEN

Department of Informatics

University of Bergen

Norway

2010

Preface

This dissertation is part of a PhD carried out at the University of Bergen at the Department of Informatics from the beginning of 2006 through to the end of 2009. The period also included a minor stay at the European Bioinformatics Institute in Hinxton, UK, and several trips abroad to international conferences and meetings. In addition to the scientific research, the PhD fellowship also included 25% teaching duties spread throughout the four years.

Serendipity

Serendipity seems to be a common feature both in life and in science. Here are three brief examples from my time in science so far. First off, I never planned to study bioinformatics. My plan was to become a biologist. But after realizing that one had to spend a lot of time doing tedious lab work and writing extensive amounts of lab reports before achieving this, I decided I would rather leave the University and become an engineer instead. However, I had to wait a semester before this plan could be set into action, and I needed something to pass the time. Somewhat randomly I ended up taking an introductory course in informatics.

Fast forward 3.5 years. I'm now doing a Master in bioinformatics, and a couple of 'summer jobs' at the Department of Informatics are advertised. I apply, resulting in the development of a tool for analyzing mass spectrometry data, and as a consequence, the original plans for my Master thesis are replaced by additional work on this mass spectrometry tool.

Fast forward again, this time around 3 years. I'm now in the middle of my PhD, still working on mass spectrometry data analysis, and an opportunity to develop a system for making mass spectrometry data publicly available comes along. And something that was originally meant to be a three month assignment, ends up as a much more extensive project, becomes an important part of my PhD, and results in a publication in Nature Biotechnology.

The bottom line seems to be that serendipity often ends up affecting even the best laid plans. And the trick seems to be to not get thrown off by this, but instead to try to make the best out of it. Given the above, this seems to have worked out pretty well for me so far.

Acknowledgements

Doing a PhD can at times be a lonely task, but it would never have been possible without the interest and support from a long list of friends and collaborators. First, I would like to thank Ingvar Eidhammer, who has been my main supervisor for both my Master and my PhD. Thank you for your continuing support and our numerous enlightening discussions. Your extensive knowledge and experience has always been a solid foundation for all our projects.

Secondly, I'd like to thank Svein-Ole Mikalsen, who has been my co-supervisor for both my Master and my PhD, and my first contact with the world of wet lab proteomics. Our discussions on such diverse topics as the minor details of mass spectrometry, to how to run Java programs, have been very motivating. And I think that both of us have learned a lot about each other's discipline's way of thinking about similar subjects.

In addition to my two official supervisors, Lennart Martens ended up becoming my unofficial co-supervisor and a central collaborator for the latter part of my PhD. I would particularly like to thank him for putting my work in a bigger context and showing me how the projects I work on constitute a small part of the bigger puzzle that is proteomics.

I would also like to thank my fellow students at the Department of Informatics and at the CBU, especially Siv Hollup and Animesh Sharma, whom I was fortunate to share an office with for a longer or shorter period of time. Our discussions (perhaps in most cases about stuff not related to our PhDs) were always entertaining and made it worthwhile going to the office.

During my PhD I was also privileged to become a part of the ProDaC (Proteomics Data Collection) community. Something which allowed me to meet a lot of interesting people and see many places I perhaps otherwise would not have visited. This has been a continual source of inspiration, and has made it a lot easier to put in some of the long hours that working on a PhD at times demands.

Finally, I would like to thank my friends and family for their continuing support. (The good-hearted jokes about my (to them) seemingly odd carrier choice have also been noted.) And last but not least, I would like to thank my parents for all their love and support throughout my life and during my academic carrier.

Abbreviations

1D, 2D, 3D	One-dimensional, Two-dimensional, Three-dimensional
CAD	Collisionally Activated Dissociation
CID	Collision Induced Dissociation
CV	Controlled Vocabulary
Da	Dalton, the atomic mass unit, named after the chemist John Dalton
de novo	From the beginning, here used to describe the process of sequencing a peptide directly from the mass spectrum, i.e., *de novo* sequencing
DNA	DeoxyriboNucleic Acid
ECD	Electron Capture Dissociation
ESI	Electro-Spray Ionization
ETD	Electron Transfer Dissociation
GUI	Graphical User Interface
HUPO	Human Proteome Organization
HUPO-PSI	HUPO Proteomics Standards Initiative
HPLC	High Pressure Liquid Chromatography
in silico	Performed on computer or via computer simulation
iTRAQ	Isobaric Tag for Relative and Absolute Quantization
LC-MS	Liquid Chromatography Mass Spectrometry
LIMS	Laboratory Information and Management System
m/z	In a mass spectrum the horizontal axis represents the mass (m) of a molecule divided by its number of charges (z), m/z is often referred to as the mass to charge ratio
MALDI	Matrix Assisted Laser Desorption and Ionization

MIAPE	Minimum Information About a Proteomics Experiment
mRNA	Messenger RiboNucleic Acid
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
ms_lims	Mass Spectrometry Laboratory Information and Management System
OLS	Ontology Lookup Service
OMSSA	the Open Mass Spectrometry Search Algorithm
ppm	parts per million
PRIDE	PRoteomics IDEntifications database
ProDaC	Proteomics Data Collection, referring to the ProDaC initiative
PSI	Proteomics Standards Initiative
PTM	Post-Translational Modification
Q-TOF	Quadrupole Time Of Flight
RNA	RiboNucleic Acid
RP	Reverse Phase
RP-HPLC	Reverse Phase High Pressure Liquid Chromatography
SAX	Strong Anion Exchange
SCX	Strong Cation Exchange
SILAC	Stable Isotope Labeling with Amino acids in Cell culture
TOF	Time Of Flight
XML	eXtensible Markup Language

Table of Contents

List of Figures

List of Tables

1 Proteomics

The topic of this thesis is proteomics. Before getting into the details of the contributions made to this field, a general introduction to the field of proteomics will be provided. The following is not intended to be a complete coverage of all areas of proteomics, but rather to serve as an overview in order to provide an understanding of the work detailed in the following chapters. For a more comprehensive overview of the field from a bioinformatics point of view see for example (Eidhammer, et al., 2007) or (Liebler, 2002), on which most of the following is based. Additional details regarding the underlying biochemistry that proteomics builds upon can be found in (Nelson and Cox, 2000) and (Creighton, 1996).

1.1 What is Proteomics?

Proteomics is one of the many 'omics' terms coined in the last couple of decades, with genomics (the study of the genomes of organisms) being among the most well known. The term proteomics is used as an analogy to genomics, based on a combination of the two terms '**prote**in' and 'gen**ome**', resulting in '**proteome**'. Proteomics can be defined as *"the study of the proteome, the protein complement of the genome, including the study of protein structure and function"* (Liebler, 2002).

The field of proteomics can roughly be divided into three central and related tasks:

- Protein Identification: Identify which protein(s) one is considering, i.e., which proteins are in a sample.
- Protein Characterization: All sorts of analyses (mainly experimental) for finding the properties of a protein. Relevant properties can be purity, charge, mass, isoelectric point, reactivity, post-translational modifications, structure, stability, amino acid composition, amino acid sequence and potential binding to other proteins.[1]
- Protein Quantification: Detecting the abundance of proteins in a sample, or across different samples which in many cases are obtained at different time points.

The most common technique to achieve these tasks is using mass spectrometry (MS), both as single MS (or simply MS) and as Tandem MS, most often referred to as

[1] Note that some of these properties can be determined by the identification of the protein, like amino acid composition and sequence.

MS/MS. But before delving into the details of mass spectrometry itself and the analysis of the results of such experiments, a closer look at the properties of the input provided to the MS instruments is necessary.

1.2 Proteins and the Proteome

"Proteins are the most abundant biological macromolecules, occurring in all cells and all parts of the cell" (Nelson and Cox, 2000). Creighton (Creighton, 1996) describes the important role of the proteins as: *"Virtually every property that characterizes a living organism is affected by proteins. Nucleic acids (…) encode genetic information – mostly specifications for the structure of proteins – and the expression of that information depends almost entirely of proteins (…)."* In other words, the proteins expressed in a given cell at a given time are essential for the properties of that cell. Zooming out, the same can also be stated for the organism as a whole.

Which proteins are expressed at a certain time, and the abundance of each individual protein, are dependent on many factors, e.g., the state of the cell and organism, resulting in an ever changing set of proteins present. This means that a cell's proteome is constantly changing. On the other hand, the proteome's genomic counterpart, the genome, is generally considered stable for a given organism (disregarding mutations etc). This calls for a closer look at the relationship between the genome and the proteome, contained in what is referred to as the central dogma in molecular biology, see Figure 1.

Figure 1: Schematic illustration of the central dogma of molecular biology, where the gene (from DNA) is transcribed into mRNA, which is then translated into a protein. Also shown is the process called replication, where DNA is replicated in order to make a copy of itself.

Note that even though mRNA can be translated to a protein, there is no guarantee that all copies of a given mRNA molecule are translated; in fact, rather the opposite is true in many cases (Collins, 2001). An mRNA molecule may also be translated many times, yielding one protein molecule per round of translation. This means that there is usually not a 1:1 ratio between the amount of mRNA produced and the abundance of the corresponding protein. In addition, mRNAs and proteins are also exposed to degradation which changes the abundance of the protein. Together this means that the ability to measure the amount of mRNA in a cell, e.g., using microarrays (Causton, et al., 2003), does not directly give you the abundance of the encoded proteins at the given time in the studied cell. To achieve this one needs to identify and quantify the proteins directly, and this is where approaches from the field of proteomics can be applied.

	Name	Abbr. 1	Abbr. 2	Mono Mass	Avg Mass	pI	Hydropathy
1	Alanine	Ala	A	71.03711	71.0788	6.01	1.8
2	Cysteine	Cys	C	103.0092	103.1448	5.07	2.5
3	Aspartate	Asp	D	115.0269	115.0886	2.77	-3.5
4	Glutamate	Glu	E	129.0426	129.1155	3.22	-3.5
5	Phenylalanine	Phe	F	147.0684	147.1766	5.48	2.8
6	Glycine	Gly	G	57.02146	57.052	5.97	-0.4
7	Histidine	His	H	137.0589	137.1412	7.59	-3.2
8	Isoleucine	Ile	I	113.0841	113.1595	6.02	4.5
9	Lysine	Lys	K	128.095	128.1742	9.74	-3.9
10	Leucine	Leu	L	113.0841	113.1595	5.98	3.8
11	Methionine	Met	M	131.0405	131.1986	5.74	1.9
12	Asparagine	Asn	N	114.0429	114.1039	5.41	-3.5
13	Proline	Pro	P	97.05276	97.1167	6.48	1.6
14	Glutamine	Gln	Q	128.0586	128.1308	5.65	-3.5
15	Arginine	Arg	R	156.1011	156.1876	10.76	-4.5
16	Serine	Ser	S	87.03203	87.0782	5.68	-0.8
17	Threonine	Thr	T	101.0477	101.1051	5.87	-0.7
18	Valine	Val	V	99.06841	99.1326	5.97	4.2
19	Tryptophan	Trp	W	186.0793	186.2133	5.89	-0.9
20	Tyrosine	Tyr	Y	163.0633	163.176	5.66	-1.3

Table 1: Listing of the 20 standard amino acids, including some of the important amino acid properties. The masses are taken from http://i-mass.com/guide/aamass.html and the other values are from (Nelson and Cox, 2000). Note that the masses are for the residues, while the pI is for the "free" amino acid.

The building blocks of proteins are the amino acids, of which there are 20 standard members, see Table 1. Each of the amino acids can be decoded from the DNA/mRNA molecules using the so-called genetic code. The genetic code is a set of rules by which information encoded in genetic material, i.e., DNA and mRNA sequences, is translated into amino acid sequences, i.e., proteins. It defines a mapping between tri-nucleotide sequences, called codons, and the amino acids, see Table 2. Four different nucleotides are used in RNA: adenine, guanine, uracil and cytosine, most often denoted as A, G, U and C. In DNA, thymine, denoted as T, is used instead of uracil.

Proteins can be studied at various levels of detail, from the amino acid sequence just described to its three-dimensional structure. Four levels are commonly used:

		2nd base			
		U	**C**	**A**	**G**
1st base	**U**	UUU Phenylalanine	UCU Serine	UAU Tyrosine	UGU Cysteine
		UUC Phenylalanine	UCC Serine	UAC Tyrosine	UGC Cysteine
		UUA Leucine	UCA Serine	UAA Stop!	UGA Stop!
		UUG Leucine	UCG Serine	UAG Stop!	UGG Tryptophan
	C	CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine
		CUC Leucine	CCC Proline	CAC Histidine	CGC Arginine
		CUA Leucine	CCA Proline	CAA Glutamine	CGA Arginine
		CUG Leucine	CCG Proline	CAG Glutamine	CGG Arginine
	A	AUU Isoleucine	ACU Threonine	AAU Asparagine	AGU Serine
		AUC Isoleucine	ACC Threonine	AAC Asparagine	AGC Serine
		AUA Isoleucine	ACA Threonine	AAA Lysine	AGA Arginine
		AUG Methionine*	ACG Threonine	AAG Lysine	AGG Arginine
	G	GUU Valine	GCU Alanine	GAU Aspartic acid	GGU Glycine
		GUC Valine	GCC Alanine	GAC Aspartic acid	GGC Glycine
		GUA Valine	GCA Alanine	GAA Glutamic acid	GGA Glycine
		GUG Valine	GCG Alanine	GAG Glutamic acid	GGG Glycine

Table 2: An overview of the standard genetic code. Color coding: yellow: non-polar; green: polar; blue: basic; red: acidic; grey: stop codon. *Note that AUG codes for both Methionine and serves as an initiation site, i.e., the first AUG in an mRNA's coding region is where the translation begins. (Figure reworked from http://en.wikipedia.org.)

- Primary Structure: The amino acid sequence.
- Secondary Structure: Regularly repeating local structures, e.g., alpha helices, beta sheets and turns.
- Tertiary Structure: The overall shape of a single protein molecule; the spatial relationship of the secondary structure elements to one another.
- Quaternary Structure: The structure formed by several protein molecules, referred to as protein subunits, functioning as a single protein complex.

This thesis will almost exclusively focus on the primary structure, from now on simply referred to as the amino acid sequence. For a more in depth discussion about the other levels see (Creighton, 1996).

1.2.1 Protein Sequence Databases

In almost all cases a protein's primary structure can uniquely identify a particular protein. As a result of this several protein sequence databases have been created in which the amino acid sequences of known proteins are accumulated. Among the most prominent are UniProtKB/Swiss-Prot (manually annotated and reviewed) and UniProtKB/TrEMBL (automatically annotated and not reviewed), both located at http://www.uniprot.org and maintained by the UniProt Consortium (UniProt Consortium, 2010). In addition to these large general databases, a multitude of other options also exist, all the way down to specialized repositories aimed at single organisms or species, e.g., the Influenza Sequence Database (http://flu.lanl.gov).

By searching such databases, either with the complete protein sequence or with parts of the sequence, it is in many cases possible to identify the protein in question.[2] This approach will be covered in more detail in Chapter 1.4.2, where identification by (partial) protein sequence information will be explained.

It is worth noting that while the protein's primary sequence is most often unique, the same is not necessarily true for a protein's accession number across different databases. As a response, both universal accession numbers based on the sequence (Babnigg and Giometti, 2006) and ways of mapping between accession numbers from different databases (Côté, et al., 2007) have been developed.

[2] Note that it is also possible to identify proteins by searching in databases of DNA sequences, but this adds additional challenges, e.g., regarding tri-nucleotide reading frames.

1.3 Mass Spectrometry

Mass spectrometry, from here onwards simply referred to as MS, has been around for a long time. However, its use as a tool for analyzing proteomics data, or more specifically proteins, is a more recent development, and this approach, often referred to as protein mass spectrometry, will be the focus of the following sections. MS can generally be defined as an analytical technique for measuring the inertial mass of (charged) molecules. In protein MS the main usage is the identification of the peptide(s) and protein(s) in a sample, which have first gone through a sample preparation stage. The main steps included in the sample preparation stage will now be sketched, followed by an overview of the properties of the most common MS instruments used in protein MS.

1.3.1 Protein Sample Preparation

The sample preparation may consist of several steps. For this brief explanation two steps will be highlighted: (i) protein separation; and (ii) protein digestion. However, it is important to note that the sample preparation includes additional steps that may influence the result of the analysis, e.g., how and how long the protein(s) are stored before being analyzed (Yi, et al., 2007). Generally it is recommended to keep all properties stable from experiment to experiment, but given that the optimal conditions can vary between experiments this may also have to be considered in the downstream analysis.

1.3.1.1 Protein Separation

In most cases it is not possible to analyze the complete proteome in a single MS experiment. This would generally result in a very complex sample that most likely would be hard to evaluate. Several procedures for separating the proteins in a sample have therefore been developed. Two-dimensional gel electrophoresis (O'Farrell, 1975) used to be the primary tool for separating proteins prior to MS analysis (Wittmann-Liebold, et al., 2006). However, in the last couple of years gel-free approaches have been developed and adopted by a growing number of labs (Gevaert, et al., 2005; Lambert, et al., 2005; Swanson and Washburn, 2005).

The common 2D gel electrophoresis first separates proteins by isoelectric focusing, followed by a 1D gel electrophoresis. Thus, the proteins are first separated along one

axis according to their isoelectric point, followed by an orthogonal separation according to their apparent molecular weight. This results in a 2D separation, with the proteins separated across the gel according to the two chosen properties. 2D gel electrophoresis has proven efficient at separating complex protein samples into discrete protein spots.[3] In addition to good separation the technique has the advantage that the spots can be visualized, either for the human eye or for further computer analysis. After 2D gel separation each spot can be excised and analyzed using MS, hopefully resulting in the identification of the protein(s) contained in each spot.

For the gel-free approaches the most common technique uses liquid chromatography (LC) to separate the proteins or peptides (see next section) prior to MS analysis. LC works by having the molecules present in a solution, which is then forced through a narrow column packed with material interacting with the molecules. The more interaction, the longer it takes the molecules to travel through the column, thus achieving separation of the molecules. Two main types of LC setups used in proteomics are: (i) reverse phase high pressure liquid chromatography (RP-HPLC), using hydrophobicity to separate the molecules; and (ii) ion exchange chromatography (either strong cation exchange (SCX) or strong anion exchange (SAX)), using charge to separate the molecules.

An alternative approach is the use of affinity chromatography, used for selective enrichment, e.g., immobilized metal ion affinity chromatography (IMAC) for phosphopeptides (Thingholm, et al., 2009), and depletion, e.g., the plasma proteome (Pernemalm, et al., 2009).

1.3.1.2 Protein Digestion

While analyzing intact proteins is possible, in most cases proteins are cleaved into smaller pieces called peptides before analysis, in a process referred to as protein digestion.[4] Cleaving a protein into peptides can be performed chemically, but is most often achieved by adding a protease to the sample mixture. The protease will in most

[3] Note that a spot may include more than one protein, and that a given protein may be spread over several spots, e.g., due to post-translational modifications.

[4] The main reason for digestion proteins into peptides is that peptides are more suitable for MS analysis, and the redundancy introduced by cleaving multiple copies of the same protein also increases the reliability of the identification. Finally, the demands on the MS instruments regarding accuracy and resolution also increases dramatically when doing the analysis on complete proteins. For further details about so-called top-down proteomics, see (Eidhammer, et al., 2007).

cases be a high-specificity protease, meaning that it cleaves the protein sequence at specific sites called cleavage sites. Trypsin is by far the most commonly used protease, but other alternatives are also employed, e.g., chymotrypsin, GluC, LysC and AspN. In addition, proteases with non-specific cleavage sites, meaning that they cleave the protein more or less randomly, are used in very specific applications. However, such proteases are not applied in the work presented in this thesis and will therefore not be covered in more detail.

Peptides obtained for MS analysis should exhibit certain characteristics in order to get optimal results. The length of the peptides should not be too short or too long, as this will interfere with their separation on an LC system, and will make them too light or too heavy to analyze accurately on a mass spectrometer. The peptides should also (for reasons that will be explained later) contain at least one amino acid residue that has the ability to accommodate a positive charge, i.e., a proton. For these reasons many potential proteases become less attractive, while trypsin often is the best choice. However, there are also situations where the properties of the proteins are better suited for a different protease. This aspect will be further explored in Chapter 2.3 and in Paper III.

The cleavage site(s) of a protease can be described as a regular expression. For example trypsin cleaves the protein sequence after the amino acids arginine (R) and lysine (K), unless followed by a proline (P), resulting in the regular expression [RK][^P] (using the Java/Perl regular expression annotation). Using these types of expressions makes it possible to *in silico* cleave a particular protein sequence with a given protease. This procedure is relied upon when matching experimentally cleaved proteins to protein sequences in a database (see Chapter 1.4.2 for more details).

1.3.2 Mass Spectrometry Instruments

Mass spectrometry instruments, or mass spectrometers, can be split into three distinct components: (i) the ion source; (ii) the mass (to charge) analyzer; and (iii) the detector. Generally the peptides start the journey in the ion source, where they are transferred to the gas phase as charged ions. These are then transported to the mass analyzer where the ions are separated according to their mass to charge (m/z) ratio. Finally, the detector records the flow of charged molecules, resulting in a mass spectrum, with the m/z value of each molecule on the horizontal axis and the intensity for each m/z on the vertical axis, see Figure 2.

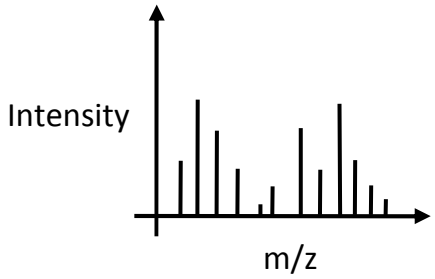

Figure 2: Schematic illustration of a mass spectrum.

In addition, most mass spectrometers used in protein MS today are capable of breaking the peptide bonds in the peptides, thus creating (peptide) fragment ions resulting in fragmentation spectra. To distinguish between the two types of spectra, the first is labeled MS spectrum, and the second MS/MS spectrum, as most instruments do this in a two-step procedure. Indeed, an MS/MS spectrum is created by selecting an m/z interval from the first MS spectrum prior to fragmentation, and a second MS analysis is then applied for the fragments produced. The analysis of both types of spectra will be discussed in more detail in Chapter 1.4.

Several types of MS instruments exist, each with their own strengths and weaknesses, and the optimal instrument may vary according to the specific problem studied or the question asked. However, some properties can be compared across experiments, mainly accuracy, precision and resolution.

1.3.2.1 Accuracy, Precision and Resolution

The three concepts of accuracy, precision and resolution are all closely related. Accuracy is here defined as the distance from the measured value to its correct mass value, and is most often given as a mass deviation, either as an absolute value, e.g., 0.5 Da, or as a relative value, e.g., 100 ppm. Relative values are used because the accuracy is usually found to be a function of the measured m/z. Precision on the other hand is the instrument's ability to reproduce the results if the experiment is repeated multiple times. Note that accuracy and precision may be improved (to the limits of the capabilities of the instrument) by proper calibration of the instrument. Finally, resolution is the instrument's ability to separate molecules with similar mass values. Resolution is most often defined using the formula: resolution = (measured

mass value) / (width of peak at a given fraction of the maximum height). The Full Width of the peak at Half its Maximum height (FWHM) is commonly used, where the width is measured at 50% of the maximum height.

Note that while the three properties are typically related there is no guarantee that an instrument with good accuracy has good precision and/or good resolution etc. In fact, different types of instruments differ quite significantly in these properties.

1.3.2.2 Ionization Methods

The transfer of the analyte molecules from a solid or liquid state to charged molecules in the gas phase is usually achieved using one of two distinct ionization methods: matrix assisted laser desorption and ionization (MALDI) or electro-spray ionization (ESI). An important distinction between these is that in ESI the molecules are dispersed as a fine aerosol, i.e., an electro-spray, into the MS instrument, thus using up the analyte over time, while in the MALDI the molecules are crystallized in a solid state on a stationary target making it possible to perform several experiments on the same sample spread over time.

1.3.2.3 Mass Analyzers

The mass analyzer separates charged molecules based on their m/z values. Several types of analyzer exist, but by far the most commonly used analyzers in protein MS are: quadrupole, quadrupole ion trap, time of flight, Fourier transform ion cyclotron resonance and orbitrap.

The quadrupole instruments consist of four circular (or ideally hyperbolic) metal rods, set perfectly parallel to each other. Quadrupoles filter the ions based on the stability of their trajectories in the oscillating electric fields applied to the rods. At a given voltage frequency, only peptides within certain m/z thresholds will be able to pass through the four rods. By varying the voltage frequency the ions can thus be separated based on their m/z values.

A variation of the quadrupole is the quadrupole ion trap. Unlike a regular quadrupole, an ion trap has the ability to contain ions within given m/z thresholds, hence the name ion trap. The trapped ions can then by targeted for further analysis, mainly for fragmentation resulting in MS/MS spectra. Actual separation of the ions by m/z

occurs when the oscillating electrical fields in the trap are tuned to eject a particular m/z range out of the trap, towards the detector.

Time of flight instruments have a different way of separating the ions. First the ions are accelerated by an electric field of known strength, resulting in equal kinetic energy for all the ions of the same charge. The velocity of the ions then depends on the mass to charge ratio of the ions, and the time required for the ion to reach a detector at a known distance is measured. The larger inertial mass of heavier ions will restrict them to lower speeds compared to lighter ions of the same charge. This difference in velocity results in different flight times, and the recorded times (together with the known experimental parameters) can be used to calculate the m/z value of each ion.

Fourier transform ion cyclotron resonance instruments determine the m/z values based on the cyclotron frequency of the ions in a very powerful, fixed electromagnetic field, and do not separate the ions in time or space. The moving charge is creating a moving electromagnetic field superimposed on the fixed electromagnetic field, which can be measured with extreme exactness. The combined cyclotron paths of the ions result in a highly complex wave, which can be translated to mass spectra by applying a Fourier transformation on this compound signal.

Orbitraps are also able to trap ions, but use a different strategy compared to the quadrupole ion trap. Here the ions are trapped using an electrostatic field. The ions orbit a central electrode, with the centrifugal forces caused by their velocity counteracting the electrostatic attraction towards this central electrode, making the ions move in complex patterns. Fourier transformations of the oscillating frequencies are then used to calculate the m/z values.

1.3.2.4 Detectors

The final component of the MS instrument is the detector, which is responsible for recording a passing or impacting ion, and forwarding this information in digital form to a computer for further processing. Most impact-based detectors rely on a form of electron cascade over multiple Faraday cups for ion impacts to translate into measurable electronic signals. These signals are in the end transformed into the MS (or MS/MS) spectrum that can be further investigated in order to identify the peptides and proteins in the analyzed sample, which is the main topic of the next chapter.

1.3.3 Peptide Fragmentation

There are several techniques for inducing peptide fragmentation, but the most common is Collision-Induced Dissociation (CID) (also referred to as Collisionally Activated Dissociation or CAD). In this approach potential energy is built up in the peptides through repeated collisions with an inert gas, e.g., argon. When an energy threshold is reached, bonds are broken and fragmentation into fragment ions and neutral losses occurs.[5] Other techniques also exist, e.g., Electron Transfer Dissociation (ETD) (Mikesh, et al., 2006) and Electron Capture Dissociation (ECD) (Zubarev, et al., 2000), but the overall concept of fragmentation remains the same. However, the types of fragment ions formed can vary depending on the technique used, see e.g., (Boersema, et al., 2009).

The fragmentation process is not yet fully understood. Although various efforts have increased the knowledge about the process, e.g., (Klammer, et al., 2008; Wysocki, et al., 2000; Zhang, 2004; Zhang, 2005), a lot still remains to be discovered. There are basically two ways in which this knowledge can be obtained: either by a bottom-up chemical approach aimed at understanding the chemical processes leading to the fragmentation, or by a top-down statistical approach where existing fragmentation data is analyzed in order to find patterns. The latter approach will be further detailed in Chapter 2.4 and in Paper IV and V.

One of the most well-known models arrived at (mainly) by using the chemical approach is the so-called Mobile Proton Model. This model states that as the dissociation energy increases, the added proton(s) will move to a protonation site, if they are not sequestered by a basic amino acid side chain (arginine, lysine or histidine). The proteins typically migrate to an atom at the amide bond, resulting in the formation of b and/or y fragment ions (see below). In addition it assumes that when the proton(s) are located at the basic amino acids one gets low proton mobilization. Further details can be found in (Paizs and Suhai, 2005).

Fragmenting peptides is not a completely random process where the peptides end up in arbitrary pieces. In most cases the peptides are mainly cleaved along the peptide backbone by cleaving the peptide bonds, which can happen in three ways. Depending on which side of the breakage the proton(s) are located six different fragment ion types can be formed from the breaking of a given peptide bond, see Figure 3. If the charge is retained on the N terminal side **a**, **b** or **c** ions are created, and if the charge

[5] In this context the peptides are more generally referred to as the precursor ions and the fragment ions as product ions.

is retained in the C terminal side **x**, **y** or **z** ions are created (Roepstorff and Fohlman, 1984).

Other types of fragmentation are also possible, e.g., internal cleavage ions (the backbone is cleaved more than once), immonium ions (a single ionized residue, formed by a combination of **a** type and **y** type cleavage) and satellite ions (ion types due to side chain cleavages). In addition, the fragments can have so-called neutral losses resulting in a mass shift of the fragment ion. Most neutral losses occur from the side chain of the amino acid residues, and consist of the loss of H_2O or NH_3, or the loss of modifications like phosphate.

1.4 Analyzing Mass Spectrometry Data

The output from mass spectrometers, i.e., the spectra, has to be analyzed by bioinformatics tools in order to identify, characterize and quantify the peptides and proteins in the samples. Before going into these details, an overview of the initial post-processing of the mass spectra will be given.

1.4.1 Raw Spectra vs. Peak Lists

Unprocessed spectra produced by MS instruments are often referred to as raw spectra, or simply raw data, and usually go through an initial post-processing step,

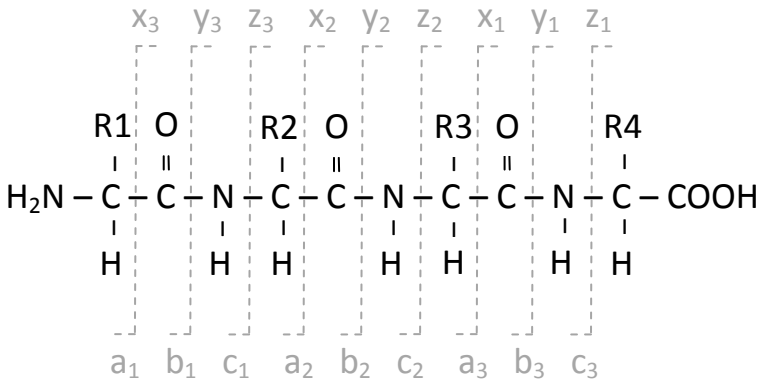

Figure 3: Illustration of the standard fragment ions which can be formed when fragmenting a peptide of length four. R1 to R4 represent the side chains of the amino acid residues.

which mainly converts the continuous mass spectrometric measurements from the raw spectra into lists of ion peaks. Post-processing detects peaks in a raw spectrum and converts them into a peak list which only contains the properties of each peak, i.e., the m/z value, the intensity, etc. This conversion from a continuous spectrum to a discrete spectrum greatly simplifies the later analysis and also reduces the space required for storing the spectra (Martens, et al., 2005).

As a part of the peak detection, the post-processing step may include one or more of the following: noise reduction, baseline correction, smoothing, intensity normalization and calibration. The process of monoisotoping or deisotoping and the removal of non-peptide masses may also be employed at this stage. For additional details see (Eidhammer, et al., 2007).

1.4.2 Protein and Peptide Identification

The identification of proteins using MS can be categorized based on the type of spectra used, i.e., (single) MS or MS/MS. When MS data from digested proteins are used the process is referred to as Peptide Mass Fingerprinting (PMF) (Cottrell, 1994). Figure 4 shows an overview of the process. Identification relying on MS/MS spectra of individual peptides is a similar process, see Figure 5, but there are important differences between the two which will now be highlighted.

1.4.2.1 Peptide Mass Fingerprinting

In peptide mass fingerprinting (PMF) the unknown protein is first cleaved into peptides, which are then inserted into an MS instrument measuring the m/z and intensity values of each peptide, resulting in an MS spectrum. Ideally, each peak in the spectrum corresponds to one peptide from the protein. This spectrum is then compared to *in silico* digested proteins from a database, and statistical methods are used to detect the best match. PMF is a fairly simple procedure, but as a result of this it has several drawbacks. For example it can only be used to identify proteins that are already in the database. Additionally, identifying more than one or two proteins at the same time becomes difficult, and the procedure is thus normally limited to highly purified proteins. Furthermore, it may be difficult to pinpoint post-translational modifications and their exact position(s) (see Chapter 1.4.3.1). This limitation will be further discussed in Chapter 2.1 and 2.3 and in Paper I and III.

MS

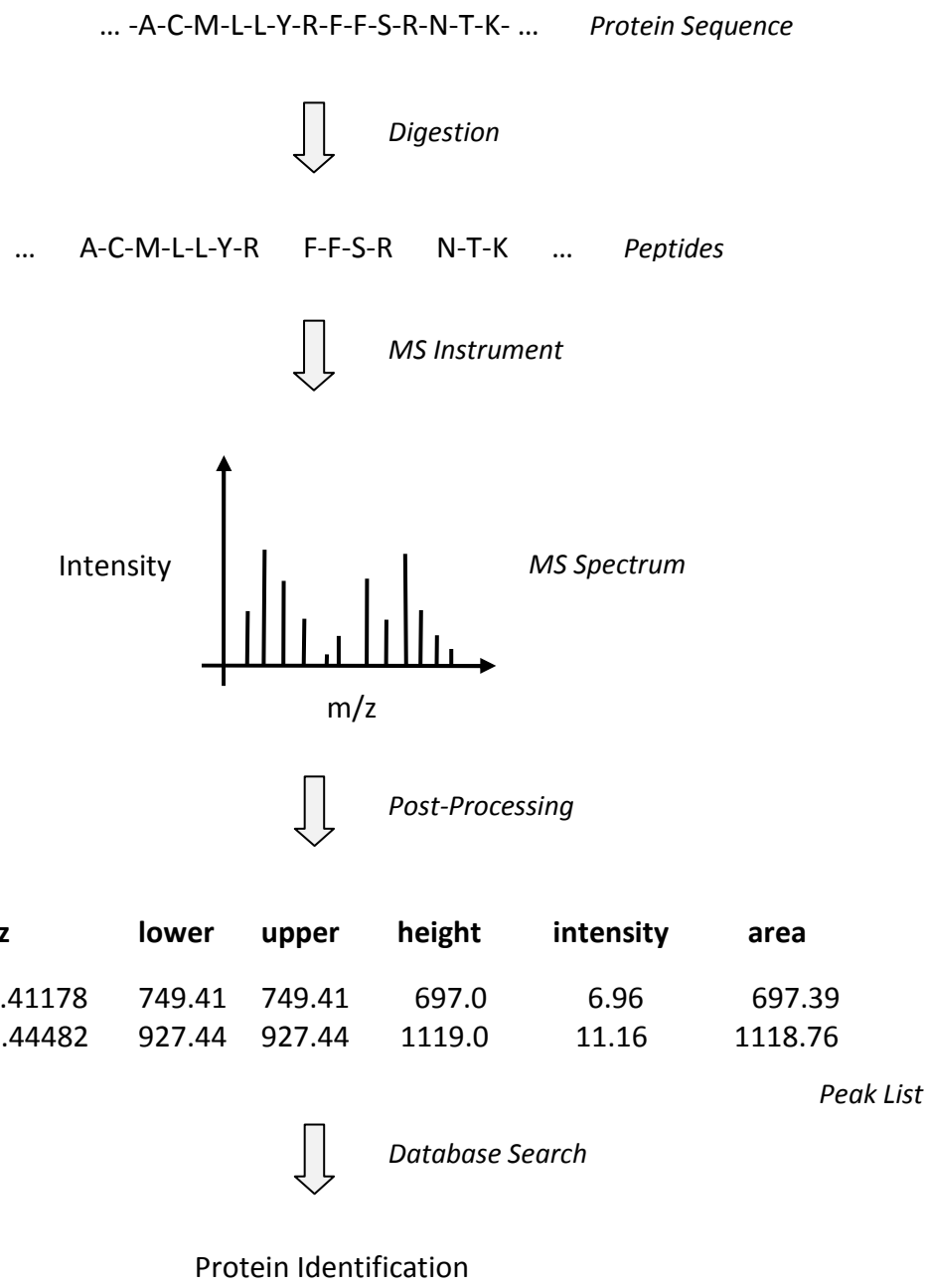

Figure 4: Simplified view of the process of protein identification via MS spectra. Ideally one peak in the peak list refers to one peak in the spectrum which again corresponds to one peptide. The additional columns shown in the peak list refer to how the discrete peak list was created based on the original continuous spectrum.

MS/MS

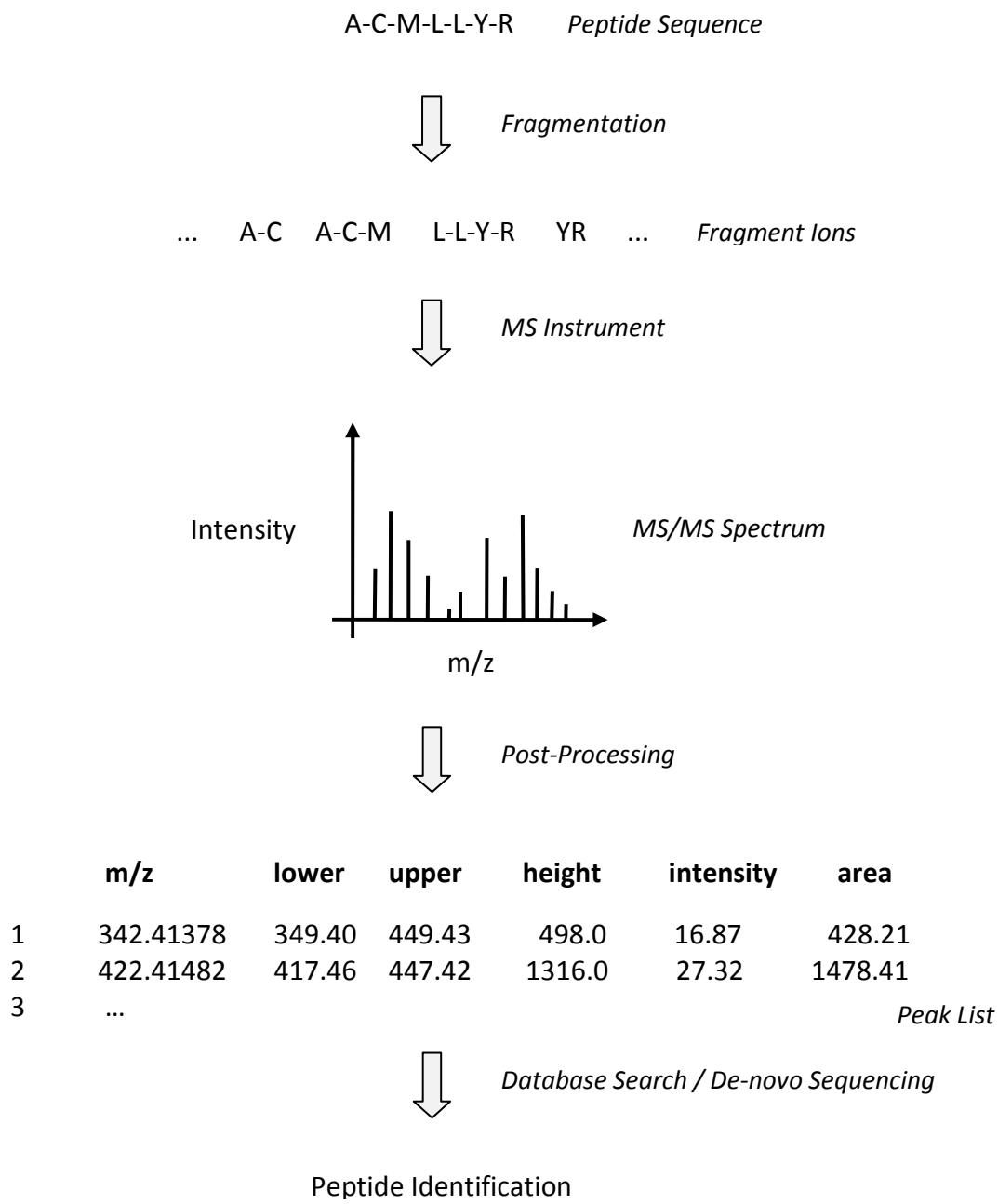

Figure 5: Simplified view of the process of peptide and protein identification via MS/MS spectra. Ideally one peak in the peak list refers to one peak in the spectrum which again corresponds to one fragment ion. The additional columns shown in the peak list refer to how the discrete peak list was created based on the original continuous spectrum.

1.4.2.2 Identification via Peptide Fragmentation

Identification of proteins via the step of peptide fragmentation resolves some of the shortcomings of PMF. In this approach the selected peptides are fragmented into fragment ions which then make up the MS/MS spectra, see Figure 5. In contrast to the spectra used in PMF, each peak in the spectrum now ideally corresponds to one fragment ion. The spectra can be identified similarly to the PMF spectra, by comparing them to *in silico* digested peptides in a database, or by a method referred to as *de novo* sequencing.

Identification of MS/MS spectra via database search is the most commonly used technique in protein MS today, and various algorithms have been developed for this purpose, e.g.,

- SEQUEST (Eng, et al., 1994; Yates, et al., 1995) [http://fields.scripps.edu/sequest]
- Mascot (Perkins, et al., 1999) [http://www.matrixscience.com]
- X!Tandem (Fenyo and Beavis, 2003) [http://www.thegpm.org/tandem]
- OMSSA (Geer, et al., 2004) [http://pubchem.ncbi.nlm.nih.gov/omssa]
- VEMS (Matthiesen, 2007) [http://personal.cicbiogune.es/rmatthiesen]

Despite the different algorithms used, three main principles can be recognized. Given an MS/MS spectrum ***S***, and a mass error threshold **ð**:

1. Find all peptides in the database having a mass similar to the precursor peptide of ***S***, within the mass error threshold **ð**.
2. Compare the theoretical fragment spectrum of each potential peptide to ***S***.
3. Calculate a score for the match between the theoretical and the experimental spectrum.

Depending on the algorithm and scoring scheme used, the possible matches are then usually ranked and a list of the best matches is presented.

If a peptide cannot be identified via a database search, for example if the peptide contains novel or unknown modifications, it is in many cases possible to extract (partial) sequence information via the method of *de novo* sequencing by utilizing the information available in the spectrum, mainly the existing peaks and the distances between the peaks. *De novo* sequencing consists of a variety of similar methods, most of which are based on graph theory, and can either be performed manually or by using automated software tools, e.g., Peaks (Ma, et al., 2003) and PepNovo (Frank

and Pevzner, 2005). A more in depth view of the details of *de novo* sequencing can be found in (Eidhammer, et al., 2007).

1.4.3 Protein Characterization

Protein characterization is usually performed after identification, in an attempt to obtain more details about a given protein, e.g., detecting the exact protein sequence (which could be altered due to mutations) or locating post-translational modifications. Such knowledge can prove essential when trying to understand a protein's role in the bigger context of a cell or an organism as a whole. However, given that these types of analyses for the most part focus on specific types or groups of proteins, and that the investigations most often are of an experimental nature, this component of proteomics, with the exception of the detection of post-translational modifications described below, will not be further detailed in this thesis.

1.4.3.1 Post-Translational Modifications

A post-translational modification is defined as a chemical modification of a protein after its translation, and can be either naturally occurring or chemically induced (either intentionally or unintentionally) during sample handling.[6] The existence of post-translational modifications is a complicating factor in all types of proteomics analysis. Generally a modification changes the mass of the modified amino acid and thus changes the mass of all peptides and fragment ions where the given residue is included.

There are in principle two ways of dealing with post-translational modifications, either by defining a set of modifications to be considered before the analysis begins, or by treating the modifications as unknown. Defining a (presumably short) list in advance results in a larger search space and in longer search times, but does not complicate the identification process significantly.[7] This is therefore the approach supported by most search algorithms. However, in many cases it is very difficult to predict all the modifications in a protein, e.g., when trying to detect novel modifications. In these situations more advanced techniques are required, but most

[6] In addition to post-translational modifications, there are co-translational modifications, i.e., modifications occurring during the translation from mRNA to protein, but at the level of mass spectrometry these two types can be handled identically.

[7] There can also be an increased chance of false positive identifications.

of these will significantly complicate the search and drastically increase the running time of the algorithm. One method that can be used to detect unknown modifications in PMF data will be discussed in Chapter 2.3 and in Paper III.

1.4.4 Protein Quantification

A task that is closely related to protein identification is the task of protein quantification, i.e., measuring the (absolute or relative) amount of protein in a sample. The ultimate goal of protein quantification is to be able to quantify the abundance of individual proteins in a sample, in many cases also across a set of samples taken at different time points. Comparing spectra obtained from different samples at different times results in a whole new set of issues regarding sample equality etc. These issues will however not be discussed further here. For the sake of simplicity it will also be assumed that only two samples are to be compared, but most approaches can be extended to more than two samples.

Protein quantification can be divided into two groups: label-based and label-free approaches. In label-based quantification either proteins or peptides from one of the samples are labeled, or both samples are labeled using different labels. Since the labels are constructed to show up as mass differences in the mass spectrometer, it becomes possible to distinguish proteins from the two samples by their mass, and extract the abundance of the detected proteins for each sample based on the measured ion intensity. Examples of label based methods are: ICAT (Gygi, et al., 1999), iTRAQ (Zieske, 2006) and SILAC (Ong, et al., 2002).

As the name suggests the label-free quantification methods do not apply any labels to the samples, but rather rely on MS data from separate LC-MS runs. Different techniques for calculating the difference between the spectra are then used to arrive at the abundance of each protein, e.g., quantification using the number of peptide identifications (spectral counting), protein sequence coverage (e.g., emPAI (Ishihama, et al., 2005)), or quantification by ion current. More details can be found in (Wong, et al., 2008).

An overview of the existing peptide and protein quantification methods, along with a discussion of issues they raise with a focus on data processing can be found in (Vaudel, et al., 2010).

1.5 Publicly Available Proteomics Data

Publicly available data repositories are the standard for most research areas in the life sciences, of which the most common examples are:

- Protein Sequences:
  - UniProtKB/Swiss-Prot and UniProtKB/TrEMBL (UniProt Consortium, 2010) [www.uniprot.org]
- Protein Structures:
  - PDB (Berman, et al., 2007) [www.rcsb.org]
- Amino Acid Modifications:
  - UniMod (Creasy and Cottrell, 2004) [www.unimod.org]
  - RESID (Garavelli, 2004) [www.ebi.ac.uk/RESID]
- Peptide and Protein Identifications:
  - PRIDE (Martens, et al., 2005) [www.ebi.ac.uk/pride]
  - PeptideAtlas (Deutsch, et al., 2008) [www.peptideatlas.org]
  - Human ProteinPedia (Keshava Prasad, et al., 2009) [www.humanproteinpedia.org]
  - GPMDB (Beavis, 2006) [http://gpmdb.thegpm.org]
- Functional Genomics / Microarray Data:
  - ArrayExpress (Brazma, et al., 2003) [www.ebi.ac.uk/microarray-as/ae]

Note that this is not an exhaustive list; it merely provides examples of some of the most commonly used repositories.

Making the data publicly available has many advantages, both at the individual data set level and perhaps most importantly at the more general repository level. Particular data sets can be tested and reanalyzed in order to verify any results published based on the data. The larger gathering of data also makes it possible to analyze all the data to look for specific patterns or properties (Klie, et al., 2008; Mueller, et al., 2008). However, the most useful feature will in many cases be the possibility of searching the data in order to identify an unknown sample and to further characterize a sample after identification, e.g., identify a protein by searching in a protein sequence database and then using information about the matching proteins to further characterize the protein in question.

Most of the data repositories mentioned above are fairly successful and already contain large amounts of data. The peptide and protein identification repositories are however lagging a bit behind compared to other data types in the life sciences. One

obvious reason for this is that in most cases the peptide and protein identification repositories have been around for a shorter period of time. But the somewhat limited success (so far) can also be explained by three complicating factors: (i) relatively complex data sets; (ii) relatively large data sets; and (iii) an (until recently) lack of data standards. For all these reasons, peptide and protein identifications present additional challenges compared to other data types. However, all of these issues are now being addressed, and the situation is starting to improve markedly. The introduction of data standards (detailed in the next sections) is the key element in solving most of the issues, but implementing ways of handling large and complex data sets in an efficient and simple manner is an equally crucial aspect.

1.5.1 Data Formats

Proteomics data have been around for a while, and it is of no surprise that a large amount of different data formats have been developed over the years. Even when limiting the scope to peptide and protein identifications from MS data alone, a long list can be produced. Here is a short list of some of the currently used data formats for MS data (both as spectra only and as spectra and identifications):

> Mascot DAT files, Mascot generic files, X!Tandem XML files, Micromass PKL files, SEQUEST DTA files, SEQUEST OUT files, OMSSA OMX files, mzXML, mzData, mzML, PRIDE XML files, Proteios XML files, VEMS PKX files, Phenyx Pidres XML files, Applied Biosystems Data Explorer PKM files, Bruker XML files, Finnigan ACS files, PerSeptive PKS files, PDF files.

It is not difficult to see that the lack of standard formats results in additional issues when the data is to be submitted to public repositories, or when data is to be transferred from one lab to another, or even inside the same lab if different instruments are used. To be able to use the data in any of these formats, the user has to be familiar with the format (in order to find the desired section of the file) and be able to extract the requested information. This puts a heavy burden on the user, which in many cases will result in potentially valuable information being disregarded due to inaccessibility.

To resolve the above situation three features have to be implemented: (i) general data standards for MS data sets; (ii) simple tools for converting data to the standard formats; and (iii) simple tools for viewing and extracting data from the standard formats. Only when all of these are in place will it be possible to shift the focus from

the data format to the actual data, which will drastically increase the usability of the available information.

1.5.2 Standards

The idea of creating standards for proteomics data is not new, and in some cases several local standards have been proposed by individual labs (McDonald, et al., 2004; Pedrioli, et al., 2004). However, it is not until recently that all of these efforts were gathered under the single umbrella of the HUPO-PSI (Human Proteome Organization – Proteomics Standards Initiative) organization, founded at the HUPO meeting in Washington in April 2002 (Kaiser, 2002). HUPO-PSI consists of various working groups focusing on different elements of the proteomics data standard: Protein Separation, Mass Spectrometry, Molecular Interactions, Protein Modifications and Proteomics Informatics. In addition, two inter-group projects are defined: Controlled Vocabularies and MIAPE (Minimum Information About a Proteomics Experiment) (Taylor, 2006). Data standards developed by the HUPO-PSI are subjected to a thorough review cycle which includes both invited experts and a period of general feedback that is open to all interested parties (Vizcaíno, et al., 2007). For more details about HUPO-PSI see http://www.psidev.info.

For the work presented in this thesis two emerging standards are particularly important: mzML and mzIdentML. mzML is a standard for mass spectrometry data, while mzIdentML is a standard aimed at capturing the different types of analyses in which MS data can be used, e.g., the identification of peptides and proteins. Both have been released in early versions, but revisions are expected in the near future. The standards are already starting to take hold in the community and the number of instruments and tools supporting these formats are increasing.

In addition to HUPO-PSI, a European 6^{th} Framework Programme funding initiative called ProDaC (Proteomics Data Collection) was also started, with the objectives to: (i) support standards development carried out by HUPO-PSI; (ii) develop conversion tools and integrate standards into products; and (iii) create a standardized workflow to submit proteomics data to central repositories. The ProDaC grant ended in March 2009 and a summary of its activities and results can be found in (Eisenacher, et al., 2009).

The adoption of the standards by the community will also be pushed forward by the scientific journals, of which a growing number are starting to demand (or at least

strongly request) the deposition of proteomics data, e.g., the raw mass spectra, in public repositories for relevant manuscripts, see for example (Editors, 2007; Editors, 2008) for the positions of the Nature Publishing Group. The number of journals enforcing this policy is already increasing, with Molecular and Cellular Proteomics recently following suit, for instance.

1.5.2.1 Controlled Vocabularies and Ontologies

Being able to read a given MS data format does not necessarily mean that one understands its contents, due to the distinction between syntax and semantics. Since a large amount of proteomics data is being produced, by a variety of different labs and by people with different backgrounds and different languages, it is not surprising that the vocabularies used to describe a given process may vary. And while these sorts of misunderstandings can be solved quite easily within a given lab, this becomes a lot more complicated in a broader, community-wide context. To solve this problem the concepts of controlled vocabularies (CVs) and ontologies were developed.

A CV is defined as a limited list of clearly defined terms, with optional relationships between the terms, while an ontology moves beyond a mere CV by attempting to extensively model a part of the real world (Martens, et al., 2008). Using CVs and ontologies makes it possible to annotate data sets in a consistent way across different labs, making it much simpler to understand an unknown data set. Annotating data using CV terms also has additional advantages, some of which will be covered in more detail in Chapter 2.5 and in Paper VI and VII.

1.5.3 Converters and Viewers

Additional tools for converting data into standard formats and for viewing or manipulating the resulting standardized data files are also necessary. Without such tools the adoption of the standards would be more difficult and occur much less rapidly in the community. An important aspect of the standardization is that it will also make it easier to submit data to public repositories, thus resulting in more publicly available data and an easier access to this data. However, for this to be possible simple tools for converting local data to the standard formats are essential. These aspects will be further covered in Chapter 2.5 and in Paper VI and VII.

2 Contributed Tools and Analyses

This chapter provides an overview of the tools and analyses contributed by the work in this thesis. For each subject the context of the given tool and/or analysis is described, and an overview of how it contributes to the field of proteomics is provided. All tools and analyses are further detailed in separate papers in the Papers section found at the end of the thesis.

2.1 Peptide Mass Fingerprinting Data Analysis

While large-scale proteomics via MS/MS is currently the most commonly used methodology in proteomics, small-scale experiments concentrating on one or a few proteins remain important as well. Such focused analyses are of particular interest when the aim is to characterize post-translational modifications in a given protein. A number of tools existed for doing small-scale protein identification, but very few of these included an administrative unit for collecting and analyzing data from several experiments on the same protein.

As a response to this we created a system called MassSorter, which is especially developed for analyzing and comparing the result of several experiments on known proteins ('known' meaning that the sequence is available and known prior to the experiments). MassSorter consists of a set of analytical tools integrated around an administrative unit that functions as a database of all performed experiments. The basis for the in-depth analysis performed by MassSorter is the comparison of the experimental and theoretical data, in many cases relying on multiple experiments for a given protein. By using the available analysis options the known protein can thereafter be analyzed for sequence coverage and different forms of modifications. In addition, unexpected cleavages can be suggested.

Note that the input to MassSorter is MS and not MS/MS data, and that this to some extent influences the abilities of the tool. The reason for not supporting MS/MS data is that the tool was aimed at (smaller) labs where MS/MS instruments were not available.

This work is further detailed in Paper I. [8]

MassSorter is available at: http://services.cbu.uib.no/software/massSorter

[8] In addition a book chapter about the tool has been written (see Chapter 4.1 for details).

2.2 Protease-Dependent Fractional Mass and Peptide Properties

Mass spectrometric analyses of peptides using protein mass fingerprints mainly rely on cleavage of proteins with proteases that have a defined specificity, and the specificities of the proteases imply that there is not a random distribution of amino acids in the peptides. This distribution had to some degree been analyzed previously for trypsin (the most common protease), but to a much lesser extent for other proteases. We therefore investigated the relationships between peptide fractional mass, pI and hydrophobicity for the three proteases trypsin, chymotrypsin and gluC, showing that the distribution of the fractional masses and the average regression lines for the fractional masses were similar, but not identical.

The analysis shows that the fractional mass and some other properties of the peptides are dependent on the protease used for generating the peptides. With the increasing accuracy of mass spectrometers it is possible to exploit the information embedded in the fractional mass of unknown peaks when analyzing peptide mass fingerprint spectra, and improving the confidence in the identifications.

This work is further detailed in Paper II.

2.3 Post-Translational Modifications and Amino Acid Substitutions

There are two main approaches for protein characterization: (i) using a predefined set of possible modifications and substitutions or (ii) performing a blind search. The first option is straightforward, but cannot (at least not directly) detect modifications or substitutions outside the predefined set. A blind search does not have this limitation, and therefore has the potential of detecting both expected and unexpected modifications and substitutions. Even previously unknown modifications can in principle be detected.

In this analysis we propose a method using blind search on protein mass fingerprinting data from two different proteases. Combining the peptide mass fingerprints from two proteases results in overlapping sequence coverage of the protein, thereby offering an alternative view of the protein and a novel way of indicating post-translational modifications and amino acid substitutions.

To show how the method can be used we also implemented a simple application called MassShiftFinder which is able to locate equal mass shifts for overlapping peptides from the two proteases used, and can indicate both post-translational modifications and amino acid substitutions. In most cases it also suggests a restricted area within the overlapping peptides where the mass shift can occur.

This work is further detailed in Paper III.

MassShiftFinder is available at: http://services.cbu.uib.no/software/massShiftFinder

2.4 Analyzing MS/MS Fragmentation Data

Ideally, evaluation of obtained peptide and protein identifications should be based on a detailed understanding of the various processes involved in acquiring the experimental data, yet for crucial steps such as fragmentation, comprehensive knowledge remains elusive. There are two main approaches by which such knowledge can be acquired, either by a chemical understanding of how the fragmentation occurs, or by a statistical analysis of available fragmentation data. Trying to utilize the large amount of protein identifications available in the ms_lims database (Helsens, et al., 2010) at Ghent University, Belgium, we implemented a tool called Fragmentation Analyzer, which makes it simple and intuitive to analyze data from MS/MS experiments in order to understand the nuances of the fragmentation process in light of experimental conditions.

The primary means through which the tool accomplishes this goal is by searching for multiple identifications of similar or equal peptides that differ in one or more user-selectable parameters, e.g., instrument type, precursor charge and post-translational modifications. The resulting information can then be used to analyze the variation in fragmentation patterns and intensities caused by using different instruments, or by post-translational modification of the peptides, amongst others.

In addition to the implementation of Fragmentation Analyzer we also used it to analyze existing peptide identifications. However, due to time limitations this work could not be completed before the writing of this thesis. An overview of our preliminary results and indications of remaining work are detailed in a separate (as yet unpublished) paper.

This work is further detailed in Paper IV and V. [9]

Fragmentation Analyzer is available at: http://fragmentation-analyzer.googlecode.com

2.5 Making Proteomics Data Sharing Easy

With the continuously growing amount of proteomics data being produced, it has become increasingly important to make these data publicly available so that they can be audited, reanalyzed and reused. In addition, more and more journals are starting to request (and even require) the deposition of MS data in publicly available repositories for submitted proteomics manuscripts. This in turn necessitates that the uploading of data to public repositories is as easy as possible, which is not always the case.

Our work focused on data deposition into the PRIDE database, which is rapidly becoming one of the most recommended data repositories for proteomics data. Several tools for submitting data to PRIDE already existed, but all of these had severe limitations, either supporting too few data formats, or being aimed almost exclusively at small scale submissions. We therefore developed PRIDE Converter, which makes it straightforward to prepare and annotate many of the most common data formats for submission to PRIDE. Through eight simple steps the relevant data are selected, annotated and converted in a wizard-like graphical user interface.

The annotation of the data using controlled vocabulary terms and ontologies makes it much easier to compare and analyze data from different sources. However, finding the correct controlled vocabulary terms can sometimes be a difficult task for the end user. As a response the Ontology Lookup Service (Côté, et al., 2006; Côté, et al., 2008) was created. For PRIDE Converter we created a user-friendly Java front end to the Ontology Lookup Service, called the OLS Dialog, which can be plugged into any application requiring the annotation of data using controlled vocabulary terms, making it possible to find and use controlled vocabulary terms without requiring any additional knowledge about web services or ontology formats.

In addition most of the data format converters had to be created more or less from scratch. Because of the scarcity of tools to access these widely used formats, some of

[9] Fragmentation Analyzer has been updated and extended since Paper IV was accepted, and now supports seven analysis types (not four as mentioned in the paper).

these were later transformed into standalone libraries that can be used to extract and visualize details from the data formats. Some of these have been published in separate publications, e.g., OMSSA Parser and XTandem Parser (see Chapter 4.1 for details).

This work is further detailed in Paper VI and VII. [10]

PRIDE Converter is available at: http://pride-converter.googlecode.com

OLS Dialog is available at: http://ols-dialog.googlecode.com

[10] Several additional papers highlighting the context of the tools or describing components of the tools have also been published (see Chapter 4.1 for details).

3 Contributing Papers

3.1 List of Included Papers

Below is a list of the papers contributing to the thesis, all of which are included at the end of the thesis, including supplementary material. All are available in the journals in which they are published, except **Paper V** which represents yet unpublished results.

Paper I:

Barsnes H, Mikalsen SO and Eidhammer I:
MassSorter: a tool for administrating and analyzing data from mass spectrometry experiments on proteins with known amino acid sequences
BMC Bioinformatics 2006 Jan 26; 7:42.
(DOI: 10.1186/1471-2105-7-42, PMID: 16438723)

Paper II:

Barsnes H, Eidhammer I, Cruciani V and Mikalsen SO:
Protease-dependent fractional mass and peptide properties
European Journal of Mass Spectrometry 2008; 14; 311-317.
(DOI: 10.1255/ejms.934, PMID: 19023148)

Paper III:

Barsnes H, Mikalsen SO and Eidhammer I:
Blind search for post-translational modifications and amino acid substitutions using peptide mass fingerprints from two proteases
BMC Research Notes 2008 Dec 19; 1:130.
(DOI: 10.1186/1756-0500-1-130, PMID: 19099572)

Paper IV:

Barsnes H, Eidhammer I and Martens L:
Fragmentation Analyzer: An open-source tool to analyze MS/MS fragmentation data
Proteomics 2010; 10(5): 1087-90.
(DOI: 10.1002/pmic.200900681, PMID: 20049869)

Paper V:

Barsnes H, Eidhammer I and Martens L:
Analyzing MS/MS Fragmentation Data
(unpublished preliminary results)

Paper VI:

Barsnes H, Vizcaino JA, Eidhammer I and Martens L:
PRIDE Converter: Making proteomics data-sharing easy
Nature Biotechnology 27, 598 - 599 (2009).
(DOI:10.1038/nbt0709-598, PMID: 19587657)

Paper VII:

Barsnes H, Côté R, Eidhammer I and Martens L:
OLS Dialog: An open-source front end to the Ontology Lookup Service
BMC Bioinformatics 2010 Jan 17; 11:34.
(DOI: 10.1186/1471-2105-11-34, PMID: 20078892)

4 Additional Work

4.1 List of Additional Papers

Following is a list of papers describing additional work performed during the period when the PhD was carried out, but which are not included in the Papers section at the end of the thesis. The list mainly consist of papers: (i) highlighting additional aspects of the tools described in the included papers; (ii) describing various contexts where the developed tools are employed; or (iii) describing various spin-off projects based on the developed tools.

Eidhammer I, Barsnes H and Mikalsen SO:
MassSorter: Peptide Mass Fingerprinting Data Analysis
Methods Mol Biol. 2008; 484: 345-59.
(DOI: 10.1007/978*-1-59745-398-1_23, PMID: 18592191)*

Barsnes H, Vizcaíno JA, Reisinger F, Eidhammer I and Martens L:
Submitting proteomics data to PRIDE using PRIDE Converter
Methods Mol Biol. 2010, Bioinformatics for Comparative Proteomics *(in press)*

Barsnes H, Huber S, Sickmann A, Eidhammer I and Martens L:
OMSSA Parser: An open-source library to parse and extract data from OMSSA MS/MS search results
Proteomics 2009; 9(14): 3772-3774.
(DOI:10.1002/pmic.200900037, PMID: 19639591)

Muth T, Vaudel M, Barsnes H, Martens L and Sickmann A:
XTandem Parser: An open-source library to parse and analyse X!Tandem MS/MS search results
Proteomics 2010; 10(7): 1522-4.
(DOI: 10.1002/pmic.200900759, PMID: 20140905)

Helsens K, Colaert N, Barsnes H, Muth T, Flikka K, Staes A, Timmerman E, Wortelkamp S, Sickmann A, Vandekerckhove J, Gevaert K and Martens L:

ms_lims, a simple yet powerful open source LIMS for mass spectrometry-driven proteomics
Proteomics 2010; 10(6): 1261-4.
(DOI: 10.1002/pmic.200900409, PMID: 20058248)

Vizcaino JA, Côté R, Reisinger F, Barsnes H, Foster JM, Rameseder J, Hermjakob H and Martens L:
The Proteomics Identifications database: 2010 update
Nucleic Acids Res. 2010 Jan;38(Database issue):D736-42
(DOI: 10.1093/nar/gkp964, PMID: 19906717)

Eisenacher M, Martens L, Hardt T, Kohl M, Barsnes H, Helsens K, Häkkinen J, Levander F, Aebersold R, Vandekerckhove J, Dunn MJ, Lisacek F, Siepen JA, Hubbard SJ, Binz PA, Blüggel M, Thiele H, Cottrell J, Meyer HE, Apweiler R and Stephan C:
Getting a grip on proteomics data – Proteomics Data Collection (ProDaC)
Proteomics 2009; 9(15): 3928-33.
(DOI:10.1002/pmic.200900247, PMID: 19637238)

Eisenacher M, Kohl M, Martens L, Barsnes H, Hardt T, Levander F, Häkkinen J, Apweiler R, Meyer HE and Stephan C:
Proteomics Data Collection – 4th ProDaC Workshop on August 15th, 2008 in Amsterdam, The Netherlands
Proteomics 2009; 9(2): 218-222.
(DOI: 10.1002/pmic.200800732, PMID: 19105180)

Eisenacher M, Martens L, Barsnes H, Hardt T, Kohl M, Häkkinen J, Apweiler R, Meyer HE and Stephan C:
Proteomics Data Collection - 5th ProDaC Workshop 4 March 2009, Kolympari, Crete, Greece
Proteomics 2009; 9(14): 3626-3629.
(DOI:10.1002/pmic.200900205, PMID: 19639582)

4.2 Web Resources

An important part in almost any tool development these days is creating a good web site for the tool. Ideally this site has to function at many levels and cater to many different types of users. First and foremost it has the important job of making the tool easily available for anybody who wants to download and use the tool. Secondly, it should also include additional information about the tool, most importantly how to use the tool: help on installing, tutorials, troubleshooting sections for common issues, how to upgrade the tool, etc.

If appropriate the site should also include information about how to reuse (parts of) the tool in other settings, e.g., how to use the OLS Dialog in other projects. Finally, as often as possible the (well-documented) source code also ought to be made available. This ensures that all the efforts that went into the making of a tool can be tapped into by other developers, making it unnecessary to reinvent the same feature whenever needed. A good example is the spectrum viewer that is used in many of the tools, from OMSSA Parser to Fragmentation Analyzer.

Having the required content is of course the key element of a web site. However, in addition the site ought to be well-structured and easy to navigate. This makes it possible for a site to contain large amounts of information tailored at different types of users without getting cluttered. Using the Google Code setup (http://code.google.com) ensures that a lot of the above points are implicitly taken care of. And with an additional effort, simple, organized and aesthetically pleasing web sites can easily be created for most tools. Google Code also supports open-source development by providing an open-access version control archive for each project.

Below is a list of web resources developed in relation to work described in this thesis. As far as possible they all try to implement the requirements just described.

Fragmentation Analyzer:	http://fragmentation-analyzer.googlecode.com
PRIDE Converter:	http://pride-converter.googlecode.com
OLS Dialog:	http://ols-dialog.googlecode.com
OMSSA Parser:	http://omssa-parser.googlecode.com
XTandemParser:	http://xtandem-parser.googlecode.com
MassShiftFinder:	http://www.bioinfo.no/software/massShiftFinder
MassSorter:	http://www.bioinfo.no/software/massSorter

5 Discussion and Future Directions

An important part of this thesis has been the development of several bioinformatics tools enabling and empowering users lacking a background in informatics to perform analyses and achieve results that would otherwise have been very difficult, e.g., converting proteomics data files with PRIDE Converter or analyzing fragmentation data with Fragmentation Analyzer. The amount of work that goes into this type of development is somewhat difficult to highlight in scientific publications, which of course mainly focus on the use of the tools and especially on the results achieved. This chapter will therefore underline the details not included in the publications, and discuss why the implementation of such tools is important for the proteomics community. Finally, some thoughts about future directions of proteomics will be given, with a focus on how these changes will affect research described in this thesis.

5.1 User Interface Design

The user interface is an essential part of any interactive program. User interfaces can be designed in many ways and there are no absolute measurements that can be used to classify one interface as better than another, and in many cases two or more very dissimilar interfaces might do the job. Different users may also prefer different interfaces. To deal with these issues, theories for designing good user interfaces have been developed. The theory presented in this section is mainly based on (Shneiderman, 1998), where eight golden rules of user interface design are described:

1. Strive for consistency.
2. Enable frequent users to use shortcuts.
3. Offer informative feedback.
4. Design dialogs to yield closure.
5. Offer prevention and simple error handling.
6. Permit easy reversal of actions.
7. Make users the initiators of actions rather than the responders to actions.
8. Reduce short-term memory load.

These underlying principles have to be interpreted, refined and extended for each environment, but if used correctly they constitute a good foundation for the development of the user interface. In addition the interface has to consider what the users want to achieve through the software and make sure that the most common features are easily accessible.

User interfaces can be divided into two groups: (i) command-based interfaces; and (ii) interfaces based on direct manipulation. In command-based interfaces the user typically communicates with the system through written commands, and all commands must comply with the syntax rules determined by the programmer. In interfaces using direct manipulation however, the user interacts with the information presented by the program, e.g., by selecting from menus, clicking on buttons or entering text in forms.

Both types of interfaces have their advantages. A command-based system is usually faster and more adaptable to advanced users' needs, but the downside is that the user has to learn and remember the syntax of the commands. Interfaces using direct manipulation are typically easier to use for new or infrequent users and the features of the program are more visible to the user, i.e., the features are presented on screen and not hidden in a possibly unknown command. A weakness of direct manipulation is that experienced/frequent users may find it slow and less adaptable to their needs. This weakness can be mitigated by adding so-called short-cuts for experienced users, resulting in a combination of the advantages of direct manipulation with the advantages of command based interfaces.

Even when a lot of effort is put into making an interface as easy to use as possible, errors will most likely still occur. There are generally two types that can occur when running a computer program: (i) program errors, i.e., errors in the source code; and (ii) user errors. Preventing errors from occurring is essential, but is not always easy. While the programmer has complete control over the source code, the same cannot be said about the user. Errors in the source code can, and should, be limited to a minimum, and most programming languages have tools for catching such errors if they do occur. The programmer can then display a message to the user explaining what went wrong. This approach can also be employed for user errors: allowing the user to make a mistake and tell him/her about it afterwards. However, a better approach is to prevent the user from making the mistake in the first place, e.g., instead of allowing the user to enter a date, make him/her select the date from a calendar. User errors can be labeled as either technical errors, e.g., clicking the wrong button by mistake, or logical errors, e.g., intentionally clicking the wrong button thinking it was correct. A good interface should try to limit both types of errors.

Another way of reducing user errors is to include help facilities and tutorials covering the main features of the tool. Instead of trial and error, the user can then peruse the help and tutorial materials before trying. This will decrease the number of errors, and if an error should occur the user has somewhere to look for an explanation. However,

the user cannot be expected to figure out the reasons for all errors. Catching and storing error messages (e.g., in a log file) is therefore important, and can substantially help reduce the time required to locate the cause of a (user-)detected bug. This feature is implemented in most of the tools described in this thesis and has proven very to be useful in practice.

Even after applying all the guidelines and theory above, a user interface may still contain issues, and an interface that is intuitive for the developer may not be intuitive for the user. All interfaces should therefore undergo thorough testing from real users and feedback from this process should be incorporated into the next release of the tool. For this to work, one should release often. As an example, see the release schedule of PRIDE Converter (see Chapter 2.5) included in the release notes found at the project's home page [http://pride-converter.googlecode.com], where new versions are released frequently, sometimes just days apart. Frequent releases are especially important in the early development stages where the main features of the tools are fine tuned. It is also important to notify users of the availability of a new version, something which is done automatically at start-up in tools like PRIDE Converter.

5.2 Enabling and Empowering Users

Almost all research performed in the field of molecular biology today to some extent relies on the use of bioinformatics, e.g., the identification of proteins using sequence databases or the image analysis of microarrays and 2D gels. The main reason for using bioinformatics is the increasing complexity and size of the obtained data sets, giving computers a significant advantage in processing time and processing accuracy compared to the human brain, in many cases even making the analyses virtually impossible without the use of computers. However, it is still essential that the users understand both the data obtained and the key details of the (bioinformatics) analysis performed, including its strengths and weaknesses. Otherwise tools may be misused and/or the results incorrectly interpreted, e.g., being able to obtain a list of protein identifications from a given sample using a search algorithm on the one hand, and understanding the properties of the identifications (regarding statistical scores etc) on the other hand is not necessarily the same thing.

One of the main advantages conveyed by good bioinformatics tools is that they enable the researchers who are closest to the data to perform the computational

analysis, rather than an external computationally oriented person, thus improving the quality of the research itself and the confidence in the achieved results. Using such tools results in more time for doing research, by reducing the time spent on manual calculations and comparisons etc. One example is the implementation of MassSorter (see Chapter 2.1) for comparing PMF data from multiple experiments. By using MassSorter this task can be performed in a matter of seconds, compared to hours or maybe days of work for manual inspection. In addition, the tool also provides better visualization of the results and supports further analyses.

Bioinformatics tools also empower researchers without local bioinformatics support to explore their data more easily. One example is the use of Fragmentation Analyzer (see Chapter 2.4) to analyze large amounts of peptide identifications in a database, an analysis that would otherwise be difficult to carry out for many labs, even when the data are easily available.

Moving the computational analysis closer to the (wet-lab) research also has additional advantages in that it makes it simpler to use the results of the analysis to design new experiments. This is especially important when doing "low-throughput" proteomics in which a smaller set of proteins (or a single protein) is analyzed for characterization purposes, e.g., to identify post-translational modifications. [11] Being able to quickly analyze the data and use the results to propose new experiments is essential in such focused settings.

5.2.1 Converting and Annotating Data

There are many formats used for storing proteomics data, and when submitting data to a repository these have to be converted to the format(s) supported by the repository. This conversion is not always straightforward and good tools are needed to simplify the process. In the following the implementation of PRIDE Converter and OLS Dialog (see Chapter 2.5) will be sketched, with a focus on how they simplify the conversion and annotation process.

A wizard-like graphical user interface was chosen for PRIDE Converter. This makes it easy to get data from the user and simplifies the process by dividing it into smaller distinct steps. Using a wizard-like graphical user interface also has the advantage that it is familiar to most users, e.g., wizards are used when installing new programs.

[11] "Low-throughput" proteomics is here used as opposed to "high-throughout" proteomics which focuses on the rapid analysis (usually identification) of a large set of peptides or proteins.

PRIDE Converter starts with the selection of the data format to be converted, followed by eight simple steps; from the selection of the files to be converted; *via* the annotation of the experiment, instrument and protocol etc; through to the setting of the output details and the actual conversion. [12] Each step handles the description/annotation of one aspect of the data, thus resulting in a simplified conversion process. One can go back and forth between the steps to alter inserted information, and multiple files can be created by repeating (some of) the steps. All user-inserted information is stored for reuse, which greatly reduces the time required for subsequent conversions. [13]

During the conversion the tool interacts with the user, most importantly for the mapping of detected post-translational modifications to their corresponding controlled vocabulary PSI-MOD modifications (Montecchi-Palazzi, et al., 2008). The most common modifications are already mapped, but the user is requested to verify the (first occurrence of a) mapping due to the mappings not necessarily being unique, e.g., C* might mean different things for different data files.

The PRIDE XML file created by the tool includes the required details from the data files, i.e., the spectra and the identifications, combined with the user input provided at the different steps, and is automatically validated and therefore ready for submission to PRIDE. Note that even though the result is an XML file, the user is not required to know anything about XML. The same is true for the data files used as input, i.e., no detailed knowledge about any of these data formats is expected either.

Being able to easily convert data files is an important improvement, but the data should also be annotated using controlled vocabulary (CV) terms (see Chapter 1.5.2.1). For this purpose the OLS Dialog was created (see Chapter 2.5). It provides a simply way of locating CV terms to annotate data, without requiring any knowledge about web services or ontology formats. Using the OLS Dialog inside PRIDE Converter provides a simple way of making sure that the data is annotated using terms that are immediately meaningful and searchable by others, thereby increasing the value of the submitted data.

PRIDE Converter and OLS Dialog have been under continuous development, and the users' informative feedback and requests for new features have been an important part of this process. As a result, PRIDE Converter has rapidly taken over as the main

[12] The steps supported do to some extent depend on the data format being converted.

[13] For some data formats, part of the information to be annotated is included in the data files, and in these cases it is automatically extracted from the data files and presented to the user for verification.

tool used for preparing files for submission to PRIDE, see Figure 6, and in most cases the files are submitted without the need for support from the PRIDE team at the European Bioinformatics Institute (EBI) [www.ebi.ac.uk/pride].

5.2.2 Analyzing Complex Data Sets

Over time a proteomics lab produces large amounts of data, which could be a valuable source of information. However, these data must therefore be stored and annotated in ways that make such further analysis possible. Using a laboratory information management system (LIMS), usually built around a relational database, is in most cases the best option. In the following the focus will be on one such system called ms_lims (Helsens, et al., 2010), and how the information contained in this database can be easily analyzed using Fragmentation Analyzer (see Chapter 2.4).

The goal of Fragmentation Analyzer is to extend and improve the understanding of

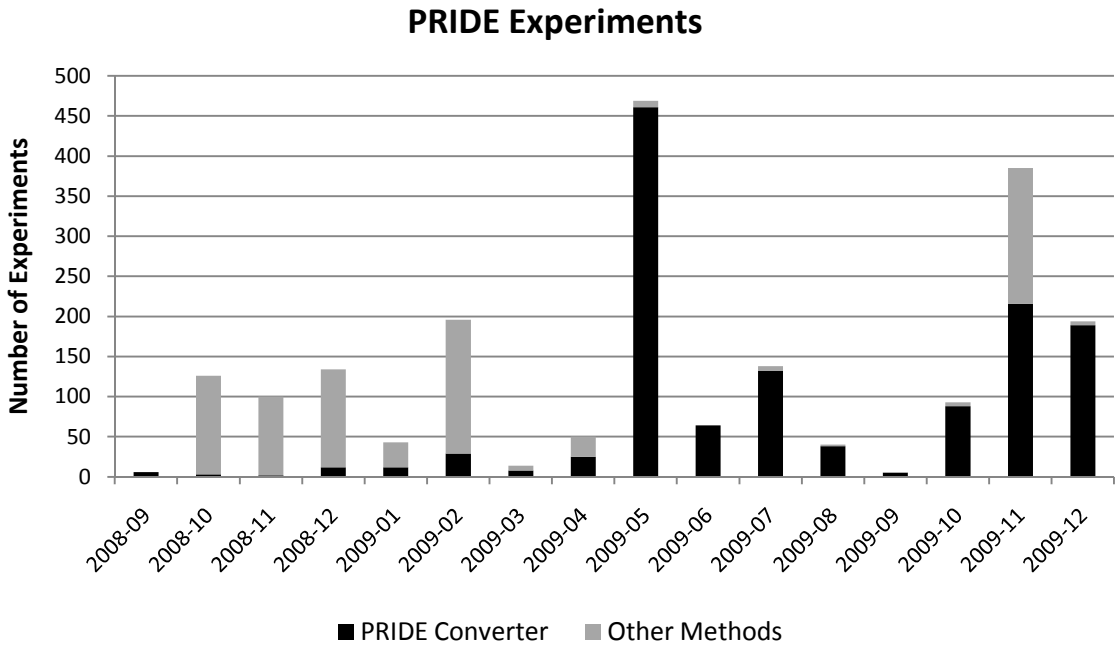

Figure 6: An overview of the number of PRIDE experiments submitted each month, from September 2008 through to December 2009. The numbers are annotated by the conversion tool used, as 'PRIDE Converter' or 'Others Methods'. Note that all the experiments submitted by other methods during November 2009 came from two projects, both submitted via MASPECTRAS (Hartler, et al., 2007) which has a built-in PRIDE export function.

the fragmentation of peptides into fragment ions, and the input consists of all the annotated fragment ions from the ms_lims database at Ghent University, Belgium, including the identified peptide sequences and the complete fragmentation spectra. In the database the desired data is spread over multiple tables, with some of these containing tens of millions of rows.

This means that just extracting the required data for analysis already presents a substantial challenge. In addition, the data comes from various instruments, the identified peptides have different precursor charges, and some of the peptides are modified and some are not. All of these categories might have to be analyzed separately, providing a further data retrieval challenge.

Fragmentation Analyzer makes it easy to extract the information from the database and to do the specialized analysis using a user-friendly graphical user interface. The tool enables the user to perform such analysis without any knowledge about the database structure or how the information is stored locally. It provides the users with easy access to the data, and with simple but efficient tools for analyzing parts of the results separately, e.g., analyze all peptide identifications from a given instrument where the peptides include a given modification. Performing such analyses is in fact quite difficult when performed by querying the database directly.

Once the desired data is obtained, the next step consists of performing the analysis. Fragmentation Analyzer supports seven analysis types, which together make it possible to obtain novel information about the data that in many cases would not have been possible to achieve from single experiments. In most cases the results are visualized in interactive plots, which the user can alter as required, e.g., by removing data series or by zooming in on interesting areas.

Fragmentation Analyzer thus enables the users to perform quite advanced analyses on complex data sets without the need for additional bioinformatics support. This makes it possible for even small labs to achieve results that would not otherwise have been possible, and serves as a good example of the importance of high-quality, user-friendly bioinformatics tools.

5.3 Open Source Software

Open access publications are becoming increasingly more common in the scientific community and the advantages of this practice are obvious; information becomes

more easily accessible, hopefully resulting in better research across the scientific community. The same principles can also be applied to open access code, often referred to as open source software. Depending on the software license used, open source software permits users to use, change, and improve the software, and to redistribute it in modified or unmodified forms. In most cases all that is needed is to include the original license in the new software, similar to science in general where existing knowledge can be built upon and reused as long as the original research is referenced.

Open source software has its challenges, perhaps chiefly the difficulty in making a profit from software that is open access, allowing everybody (at least in theory) to build and distribute their own versions. In many cases this problem is solved by relying on a business model where the code may be freely available, but users are instead charged for support and consultancy. For most software developed in research projects however, making a profit is never the main goal and in most cases the achievable profit is insignificant anyway.

One example of open source software in the scientific community is the implementation of various statistical analysis methods, which is needed in most types of tools analyzing bioinformatics data. Instead of every project implementing its own version of the required analysis type, the analysis ought to be implemented in general open source libraries. Many such libraries have been created, e.g., R [http://www.r-project.org] and JFreeChart [http://www.jgoodies.com], which dramatically reduces the work required to implement and use a large variety of statistical methods and plotting the results of the analysis.

Making the source code publicly available also has the benefit that it becomes much easier for others to participate in the development of the tool, and instead of getting a bug report the person detecting a bug could ideally also fix the bug. Adding new components or features and making the tool work in new settings also becomes much simpler when the source code is available.

However, making a project open source requires an additional effort from the developers. The code ought to be well-structured, well-documented, easy to maintain and easy to understand by others. In some cases this non-trivial additional workload might result in projects not becoming open source, so a stronger incentive is sometimes needed. Some scientific journals recommend that projects are made open source when publishing, but very few are demanding open source. Given the additional benefits that open source projects provide (compared to closed-source

projects), this recommendation should be implemented by all journals publishing manuscripts describing tools, algorithms and methods, and perhaps changed from a recommendation to a strong request. This would result in more open source projects being published and each paper would contribute more to the community. In addition, this would reduce the problem of the variable lifespan of academic projects, which in many cases are left to die as soon as the people responsible move on to new projects or new jobs. Open source projects can live on even long after the original developers have left the project, thereby increasing the value of the developed tools beyond their immediate purpose.

Open source software can to some extent be compared to the publishing of proteomics data. While publishing protein/peptide identifications alone can be useful, a lot more information is included when also publishing the mass spectra. Similarly for software, while publishing the results of the software and the software itself is valuable, the value increases substantially if in addition the source is made available.

5.4 Future Directions

In the world of science nothing is written in stone and future developments will most likely change many of our conceptions on how proteomics ought to be done, and new and possibly surprising pieces in the bigger puzzle of molecular biology will almost certainly be found. This section outlines some relevant future directions of proteomics, and discusses how this may influence the field.

5.4.1 Standardized and Open Access Proteomics

The work of standardizing proteomics, and in addition making sure that proteomics data are made publicly available, is perhaps the most important topic in proteomics today. As soon as standards for proteomics data have been implemented in all the instrument software and tools, and are in everyday use by researchers, the field of proteomics will change considerably. This will of course not happen overnight, but a widespread adoption of the standards within a couple of years should not be considered overly optimistic. In this setting, much of the work today spent on converting data or trying to understand an unknown data format can rather be spent doing actual research.

Mandatory submission of data to public repositories will also become much simpler, given that the data does not have to be converted as this will already be a part of the process in creating the data files. As a consequence the amount of data in public repositories will increase and enable the extraction of new knowledge from the accumulated data, which otherwise would have been out of reach. To make this possible the data have to be annotated by controlled vocabulary terms, and this is where tools such as the OLS Dialog can be utilized, see Chapter 2.5.

Standardization will also free up a lot of programming time currently spent on repeating similar types of analysis for different instruments or data types. Instead the efforts can be directed at research-oriented tools trying to extract statistics-based knowledge from all available data. One example here would be a connection between the Fragmentation Analyzer and the PRIDE database, making it possible to analyze all data in PRIDE simultaneously. This could for example be used to analyze differences due to using different search algorithms, instruments or protocols, or to compare spectra and identifications to assess false positives.

5.4.2 Improved Protein Quantification

The field of protein quantification is rapidly maturing and the techniques employed are constantly improving. In the not too distant future protein quantification will hopefully become a technique that can be relied upon and trusted with even more confidence. This will open up a whole new vista of opportunities for innovative and interesting research. Improved techniques for protein quantification can also be connected with one of the most prominent buzz words in molecular biology today: biomarkers. Biomarkers are defined as characteristic biological properties that can be detected and measured, most often in blood, urine or tissue, to indicate something about the state of the cell or organism, e.g., if the individual is affected by a certain disease or not (Rosner, 2009).

While biomarkers have been a catchphrase in the research community for some time the applicable results of the research are lagging behind and very few biomarkers have made it past the pre-clinical test stage, see for example (Taube, 2009). Improved techniques for measuring and quantifying proteins might just be the break-through that the field is waiting for.

5.4.3 Integrative Omics Research

The ultimate goal of biological research is of course to put everything together and reveal the bigger picture. This cannot be achieved by proteomics alone. Only by combining and integrating results from many of the distinct omics fields can this be accomplished. For example by combining genome data with transcriptomes [14] (from microarrays or next-generation sequencing) and quantitative proteomics data, novel and exciting results might be obtained. Most of these results would not have been possible by using just one of the data types. Examples of integrative omics approaches are given in (Thongboonkerd, 2005) and (Fukushima, et al., 2009).

Effective integrative omics research relies on a high level of maturity in all the relevant omics fields. The standardization activities that proteomics is currently undergoing are thus an essential contribution in preparing the field for the next generation integrative omics research. Exciting new opportunities for research will therefore certainly become available in the foreseeable future, through hard work and perhaps also a little serendipity.

[14] The transcriptome is the set of all RNA molecules, including mRNA, rRNA, tRNA, and non-coding RNA produced in one or a population of cells.

6 References

Babnigg, G. and Giometti, C.S. (2006) A database of unique protein sequence identifiers for proteome studies, *Proteomics*, **6**, 4514-4522.

Beavis, R.C. (2006) Using the global proteome machine for protein identification, *Methods Mol Biol*, **328**, 217-228.

Berman, H., Henrick, K., Nakamura, H. and Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data, *Nucleic Acids Res*, **35**.

Boersema, P.J., Mohammed, S. and Heck, A.J. (2009) Phosphopeptide fragmentation and analysis by mass spectrometry, *J Mass Spectrom*, **44**, 861-878.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P. and Sansone, S.A. (2003) ArrayExpress--a public repository for microarray gene expression data at the EBI, *Nucleic Acids Res*, **31**, 68-71.

Causton, H.C., Quackenbush, J. and Brazma, A. (2003) *Micoarray - Gene Expression Data Analysis*. Blackwell Publishing.

Collins, F.S. (2001) Contemplating the end of the beginning, *Genome Res*, **11**, 641-643.

Côté, R.G., Jones, P., Apweiler, R. and Hermjakob, H. (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries, *BMC Bioinformatics*, **7**.

Côté, R.G., Jones, P., Martens, L., Apweiler, R. and Hermjakob, H. (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries, *Nucleic Acids Res*, **36**.

Côté, R.G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R. and Hermjakob, H. (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases, *BMC Bioinformatics*, **8**.

Cottrell, J.S. (1994) Protein identification by peptide mass fingerprinting, *Pept Res*, **7**, 115-124.

Creasy, D.M. and Cottrell, J.S. (2004) UniMod: Protein modifications for mass spectrometry, *Proteomics*, **4**, 1534 - 1536.

Creighton, T.E. (1996) *Proteins - Structure and Molecular Properties*. W.H. Freeman and Company.

Deutsch, E.W., Lam, H. and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows, *EMBO Rep*, **9**, 429-434.

Editors (2007) Democratizing proteomics data, *Nat Biotechnol*, **25**, 262.

Editors (2008) Thou shalt share your data, *Nat Methods*, **5**, 209-209.

Eidhammer, I., Flikka, K., Mikalsen, S.O. and Martens, L. (2007) *Computational Methods for Mass Spectrometry Proteomics*. John Wiley & Sons, Ltd.

Eisenacher, M., Martens, L., Hardt, T., Kohl, M., Barsnes, H., Helsens, K., Häkkinen, J., Levander, F., Aebersold, R., Vandekerckhove, J., Dunn, M.J., Lisacek, F., Siepen, J.A., Hubbard, S.J., Binz, P.A., Blüggel, M., Thiele, H., Cottrell, J., Meyer, H.E., Apweiler, R. and Stephan, C. (2009) Getting a grip on proteomics data - Proteomics Data Collection (ProDaC), *Proteomics*, **9**, 3928-3933.

Eng, J., McCormack, A.L. and Yates, J.R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *J Am Soc Mass Spectrom*, **5**, 976–989.

Fenyo, D. and Beavis, R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes, *Anal Chem*, **75**, 768-774.

Frank, A. and Pevzner, P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling, *Anal Chem*, **77**, 964-973.

Fukushima, A., Kusano, M., Redestig, H., Arita, M. and Saito, K. (2009) Integrated omics approaches in plant systems biology, *Curr Opin Chem Biol*, **13**, 532-538.

Garavelli, J.S. (2004) The RESID Database of Protein Modifications as a resource and annotation tool, *Proteomics*, **4**, 1527-1533.

Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W. and Bryant, S.H. (2004) Open mass spectrometry search algorithm, *J Proteome Res*, **3**, 958-964.

Gevaert, K., Van Damme, P., Martens, L. and Vandekerckhove, J. (2005) Diagonal reverse-phase chromatography applications in peptide-centric proteomics: ahead of catalogue-omics?, *Anal Biochem*, **345**, 18-29.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H. and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, *Nat Biotechnol*, **17**, 994-999.

Hartler, J., Thallinger, G.G., Stocker, G., Sturn, A., Burkard, T.R., Körner, E., Rader, R., Schmidt, A., Mechtler, K. and Trajanoski, Z. (2007) MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data, *BMC Bioinformatics*, **8**, 197.

Helsens, K., Colaert, N., Barsnes, H., Muth, T., Flikka, K., Staes, A., Timmerman, E., Wortelkamp, S., Sickmann, A., Vandekerckhove, J., Gevaert, K. and Martens, L. (2010) ms_lims, a simple yet powerful open source LIMS for mass spectrometry-driven proteomics, *Proteomics*, **10**, 1261-1264.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J. and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein, *Mol Cell Proteomics*, **4**, 1265-1272.

Kaiser, J. (2002) Proteomics. Public-private group maps out initiatives, *Science*, **296**, 827.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., S., B., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. and Pandey, A. (2009) Human Protein Reference Database--2009 update, *Nucleic Acids Res*, **37**, D767-772.

Klammer, A.A., Reynolds, S.M., Bilmes, J.A., MacCoss, M.J. and Noble, W.S. (2008) Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification, *Bioinformatics*, **24**, i348-356.

Klie, S., Martens, L., Vizcaíno, J.A., Côté, R., Jones, P., Apweiler, R., Hinneburg, A. and Hermjakob, H. (2008) Analyzing large-scale proteomics projects with latent semantic indexing, *J Proteome Res*, **7**, 182-191.

Lambert, J.P., Ethier, M., Smith, J.C. and Figeys, D. (2005) Proteomics: from gel based to gel free, *Anal Chem*, **77**, 3771-3787.

Liebler, D.C. (2002) *Introduction to Proteomics - Tools for the New Biology*. Humana Press.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. and Lajoie, G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry, *Rapid Commun Mass Spectrom*, **17**, 2337-2342.

Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J. and Apweiler, R. (2005) PRIDE: the proteomics identifications database, *Proteomics*, **5**, 3537-3545.

Martens, L., Nesvizhskii, A.I., Hermjakob, H., Adamski, M., Omenn, G.S., Vandekerckhove, J. and Gevaert, K. (2005) Do we want our data raw? Including binary mass spectrometry data in public proteomics data repositories, *Proteomics*, **5**, 3501-3505.

Martens, L., Palazzi, L.M. and Hermjakob, H. (2008) Data standards and controlled vocabularies for proteomics, *Methods Mol Biol*, **484**, 279-286.

Matthiesen, R. (2007) Virtual Expert Mass Spectrometrist v3.0: an integrated tool for proteome analysis, *Methods Mol Biol*, **2007**, 121-138.

McDonald, W.H., Tabb, D.L., Sadygov, R.G., MacCoss, M.J., Venable, J., Graumann, J., Johnson, J.R., Cociorva, D. and Yates, J.R., III (2004) MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications, *Rapid Commun Mass Spectrom*, **18**, 2162-2168.

Mikesh, L.M., Ueberheide, B., Chi, A., Coon, J.J., Syka, J.E., Shabanowitz, J. and Hunt, D.F. (2006) The utility of ETD mass spectrometry in proteomic analysis, *Biochim Biophys Acta*, **1764**, 1811-1822.

Montecchi-Palazzi, L., Beavis, R., Binz, P.A., R.J., C., Cottrell, J., Creasy, D., Shofstahl, J., Seymour, S.L. and Garavelli, J.S. (2008) The PSI-MOD community standard for representation of protein modification data, *Nat Biotechnol*, **26**, 864-866.

Mueller, M., Vizcaíno, J.A., Jones, P., Côté, R., Thorneycroft, D., Apweiler, R., Hermjakob, H. and Martens, L. (2008) Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets, *Proteomics*, **8**, 1138-1148.

Nelson, D.L. and Cox, M.M. (2000) *Lehninger: Principles of Biochemistry*. Worth Publishers.

O'Farrell, P.H. (1975) High resolution two-dimensional electrophoresis of proteins, *J Biol Chem*, **250**, 4007-4021.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A. and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC,

as a simple and accurate approach to expression proteomics, *Mol Cell Proteomics*, **1**, 376-386.

Paizs, B. and Suhai, S. (2005) Fragmentation pathways of protonated peptides, *Mass Spectrom Rev*, **24**, 508-548.

Pedrioli, P.G., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R., Cheung, K., Costello, C.E., Hermjakob, H., Huang, S., Julian, R.K., Kapp, E., McComb, M.E., Oliver, S.G., Omenn, G., Paton, N.W., Simpson, R., Smith, R., Taylor, C.F., Zhu, W. and Aebersold, R. (2004) A common open representation of mass spectrometry data and its application to proteomics research, *Nat Biotechnol*, **22**, 1459-1466.

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, **20**, 3551-3567.

Pernemalm, M., Lewensohn, R. and Lehtiö, J. (2009) Affinity prefractionation for MS-based plasma proteomics, *Proteomics*, **9**, 1420-1427.

Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides, *Biomed Mass Spectrom*, **11**, 601.

Rosner, M.H. (2009) Urinary biomarkers for the detection of renal injury, *Adv Clin Chem*, **49**, 73-97.

Shneiderman, B. (1998) *Designing the User Interface*. Addison Wesley.

Swanson, S.K. and Washburn, M.P. (2005) The continuing evolution of shotgun proteomics, *Drug Discov Today*, **10**, 719-725.

Taube, S.E. (2009) Biomarkers in oncology: trials and tribulations, *Ann N Y Acad Sci*, **1180**, 111-118.

Taylor, C.F. (2006) Minimum reporting requirements for proteomics: a MIAPE primer, *Proteomics*, **6 Suppl 2**, 39-44.

Thingholm, T.E., Jensen, O.N. and Larsen, M.R. (2009) Enrichment and separation of mono- and multiply phosphorylated peptides using sequential elution from IMAC prior to mass spectrometric analysis, *Methods Mol Biol*, **527**, 67-78.

Thongboonkerd, V. (2005) Genomics, proteomics and integrative "omics" in hypertension research, *Curr Opin Nephrol Hypertens*, **14**, 133-139.

UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010, *Nucleic Acids Res*, **38**, D142-148.

Vaudel, M., Sickmann, A. and Martens, L. (2010) Peptide and protein quantification: a map of the minefield, *Proteomics*, **10**, 650-670.

Vizcaíno, J.A., Martens, L., Hermjakob, H., Julian, R.K. and Paton, N.W. (2007) The PSI formal document process and its implementation on the PSI website, *Proteomics*, **7**, 2355-2357.

Wittmann-Liebold, B., Graack, H.R. and Pohl, T. (2006) Two-dimensional gel electrophoresis as tool for proteomics studies in combination with protein identification by mass spectrometry, *Proteomics*, **6**, 4688-4703.

Wong, J.W., Sullivan, M.J. and Cagney, G. (2008) Computational methods for the comparative quantification of proteins in label-free LCn-MS experiments, *Brief Bioinform*, **9**, 156-165.

Wysocki, V.H., Tsaprailis, G., Smith, L.L. and Breci, L.A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation, *J Mass Spectrom*, **35**, 1399-1406.

Yates, J.R., III, Eng, J.K., McCormack, A.L. and Schieltz, D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database, *Anal Chem*, **67**, 1426-1436.

Yi, J., Kim, C. and Gelfand, C.A. (2007) Inhibition of intrinsic proteolytic activities moderates preanalytical variability and instability of human plasma, *J Proteome Res*, **6**, 1768-1781.

Zhang, Z. (2004) Prediction of low-energy collision-induced dissociation spectra of peptides, *Anal Chem*, **76**, 3908-3922.

Zhang, Z. (2005) Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges, *Anal Chem*, **77**, 6364-6373.

Zieske, L. (2006) A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies, *J Exp Bot*, **57**, 1501-1508.

Zubarev, R.A., Horn, D.M., Fridriksson, E.K., Kelleher, N.L., Kruger, N.A., Lewis, M.A., Carpenter, B.K. and McLafferty, F.W. (2000) Electron capture dissociation for structural characterization of multiply charged protein cations, *Anal Chem*, **72**, 563-573.